NATIONAL SCIENCE FOUNDATION
Information Technology Research (ITR) Program
Technical/Programmatic Project Goals and Objectives

Building the Tree of Life – A National Resource for Phyloinformatics and Computational Phylogenetics

1. The Awardees shall use reasonable efforts to achieve the following proposed goals and objectives:

Year 1

- Design the platform: main activities will include software architecture, database design, hardware acquisition and integration.
- Define a system architecture (hardware and software), with a small collection of simple APIs.
- Order the hardware and install much of it.
- Have wrappers around existing software from team members (e.g., PAUP*, MrBayes, Mesquite, GRAPPA, etc.) to enable their temporary use on systems.
- Define a "Version 1" schema for the integration of TreeBase with other data resources and produce a plan for federating the resulting database within the larger context of SDSC's unified bioinformatics resources.
- Start transferring data from TreeBase into database system.
- Produce a first-generation set of fairly modest benchmarks and develop a plan for producing new, more complex models of evolution.
- Develop initial web site aimed mostly at researchers in the area.

Year 2

- Implement the architecture designed in the first year. This will provide the framework within which software modules (for phylogeny reconstruction, post-tree analyses, performance evaluation, and simulations of evolution) can be placed.
- Start populating the framework with solutions modules.
- Have a rough user interface in place for temporary use, with partial integration of the database, solution modules, and current simulation tools.
- Implement and partially test new models of evolution
- Make available for testing an alpha release of the software suite.
- Have new entrances on the project web for college students from Biology and from Computer Science.

Year 3

- Populate the framework implemented in the second year. This includes populating the database with biological data and large simulated datasets as well as populating the computational framework with reconstruction algorithms, evaluation modules, etc.
- Begin beta testing by outside users (mostly collaborators, but also other ATOL investigators, students, and participants in the annual workshops).
- Replace most early, re-wrapped software (PAUP*, Mr Bayes, etc.) with new modules produced by team members.

- Ensure user Interface roughly at the level of 2003-stock Mesquite and smoothly integrate database functions with the various solution modules.
- Undergo testing of New algorithmic developments to handle scaling to very large datasets.
- Identify specific simulations to produce the very large datasets to be used in the next phase.
- Make available software suite in beta release and run on laptops as well as on small SMPs.
- Feature a prototype educational interface and public interface on the website.

Year 4

- Use the results of simulation research to produce biologically realistic datasets containing a few million sequences
- Use the (much smaller) biological datasets provided by the ATOL projects to help these projects analyze their data for publication.
- Conduct extensive tests of scalability using very large simulated datasets and make recommendations for analyses.
- ATOL partners will use software in earnest.
- Refine the user interface through testing in workshops and collaborations.
- Ensure the software suite is usable on any platform and runs well on large SMP clusters.
- Have entrances for children, adults, pupils, students, and researchers on the website
- Link website closely with the Tree of Life web site of David Maddison.
- Produce educational modules for high-school teachers.
- Carefully review accomplishments and projected results against success criteria and plans formulated for acquiring funds to ensure continuing support for the resource.

Year 5

- Include in framework modules contributed by international collaborators and others.
- Include in the database a large variety of benchmarks, as well as many datasets from ATOL partners.
- On simulated data, produce accurate reconstruction for datasets of one million items.
- Conduct successful analyses of large biological datasets with ATOL partners.
- Lay the groundwork needed to enable software suite for Grid usage.
- Release a final version of Web interface, including appropriate mirroring facilities.
- Have a traveling exhibit in place with the Peabody museum
- Transition to a new funding model and complete a new mandate in consultation with funding organizations
- Make adjustments to personnel and objectives.

2. Detailed Times Lines

Time lines are given year by year, broken down in each by focus group. The focus groups are algorithms research, simulation research, database research, software design, outreach, and the professional team at SDSC.

- The algorithms group includes Warnow (coordinator) and Hunt at UT, Rao, Karp, Papadimitriou, Russell, and Myers at Berkeley, Moret, Bader, andWilliams at UNM, and Huelsenbeck at UCSD.
- The simulation group includes Kim (coordinator) and Kannan at UPenn, Hillis and Meyers at UT, Muse at NCSU, and Turner at Yale, plus, on the evaluation side, Moret at UNM and Warnow at UT.
- The software design group includes Swofford (coordinator) at FSU, W. Maddison (coordinator) at UBC, D. Maddison at U.Az, Lewis at UConn, Wheeler at AMNH, Warnow at UT, Moret, Bader, and Williams at UNM, and Miller, Berman, and Bourne at SDSC.
- The database group includes Miranker (coordinator) at UT, Donoghue at Yale, Mishler and Papadimitriou at Berkeley, Piel at SUNY, and Bourne at UCSD, plus a member of the professional staff at SDSC.
- The outreach group includes Donoghue (coordinator) at Yale, Mishler (coordinator) at Berkeley, Wheeler at AMNH, Warnow at UT, D. Maddison at U.Az, and Miller at SDSC.
- The professional group at SDSC is headed by Miller, in constant consultation with the PIs and various focus leaders.

Year 1

Algorithms: The project's algorithm researchers naturally form two closely collaborating groups. UT and UNM collaborate extensively on gene-order phylogeny, reticulate evolution, and large-scale phylogeny reconstruction, and have several collaborative ITR grants in these topics. Berkeley and UCSD share interests in MCMC methods (Russell and Huelsenbeck), and also have a large number of theoretical computer science researchers (Karp, Papadimitriou, Rao, and Myers). The two groups will collaborate on all algorithms problems, but interactions within a group are likely to be more frequent and focused. During the first year, the Berkeley group will develop lower bounds and approximation algorithms for maximum parsimony, and will also study theoretical approaches to some novel difficult computational problems arising from reconstructing phylogenies, including gene family evolution and genomic alignment. The Berkeley group (esp. Russell) will also begin to collaborate with UCSD (Huelsenbeck) on testing scalability of MCMC methods. Members of the UNM and UT group (Moret, Bader, Williams, Warnow, Linder, and Jansen) will continue their collaboration on developing algorithms for inferring phylogenies from gene order and content, detection of reticulation and reconstruction of reticulate phylogenies, and using divide-and-conquer methods for finding better solutions to maximum parsimony and maximum likelihood. Hunt (at UT) will examine the use of symbolic representation in speeding up phylogenetic analyses.

Simulations: Simulations researchers will (1) curate phylogenetically relevant key data from molecular databases (in collaboration with ATOL-Sanderson team), (2) statistically characterize key molecules (Muse, Hillis), (3) develop data management strategies for simulated datasets of several million branches (Kim, in collaboration with Miranker and in consultation with Penn computer scientist Susan Davidson), (4) develop computational strategies for scalable simulation (Kim, Kannan, Moret, Warnow), and (5) develop models of molecular evolution for key molecules (Muse, Hillis). Two workshops are planned in the first year: one focusing on complex genome simulations (Meyers), and one focusing on phylogenetic simulations (Kim). Project biologists Linder and Jansen will be involved in the workshops involving phylogenetic simulations.

Software Design: Our principal software designers, Swofford and W. Maddison, will make frequent visits to SDSC, to confer with the professional team in designing the overall system architecture and, in particular, the initial set of APIs for the interface software. The phorest subteam (Swofford, Lewis, and Holder) will proceed with continued implementation of its original (and already started) project, using lessons from its design to prepare a better one for our project and modifying their work as needed to ensure that it will fit easily within the IT resource a year down the line. Working with the database group, the software design group will develop standards for data input and exchange formats, drawing from experience with Nexus and other existing methods, incorporating XML and possibly other emerging metalanguages. Moret, Bader, Berman, and Wheeler will consult with the group on issues related to algorithm engineering and high-performance computing; in particular, they will set up a "stunt run:" using NCAR allocations, they will run a prototype analysis of a few million-sequence datasets (produced with a first, rough simulation tool), in order to demonstrate feasibility and gather some preliminary computational data.

Databases: Miranker, Donoghue, Piel, and Mishler will consult frequently with Bourne and the SDSC team to choose tools, design the initial DB schema, figure out how best to leverage existing software tools developed at SDSC for PDB and other DB projects, and devise a plan to transfer the contents of TreeBase to the IT resource. Piel, as the main person in charge of the transfer and of later curation, will spend significant time at SDSC itself. Miranker will initiate research in novel DB structures to support complex queries on collections of structured objects such as trees.

Outreach: Mishler will start immediately on the development of weekend workshops at Berkeley with the help of a part-time assistant and begin the series in the second half of the year. The SDSC team will set up the first web site. The program for minority undergraduates from Lehman College will recruit its first participants for Summer 2004. The simulation team will organize community workshops for evolutionary biologists and others working in modeling evolution. Mutual visits with ATOL research teams and with Doolittle's group in Canada will be initiated.

SDSC: The SDSC team will establish a scope of work document specifying required features for the database, the user interface, the computational support, and the modules to be integrated in later years. It will develop a first version of a web interface and of a database schema, identify selected software from existing team software and mount it on the resource under suitable wrapping interfaces. It will set up a collection of internal web pages for self-documentation, recording the team's plans for module design, implementation, integration, and documentation. It will negotiate with vendors, identify appropriate hardware, acquire it, and install it.

Year 2

Algorithms: Algorithms researchers will develop novel approaches for these problems, and begin to experimentally evaluate the most promising of the approaches. In addition, they will continue to develop new and improve existing algorithms based upon insights that are gained from the experiments. The Berkeley and UCSD faculty will focus on developing new MCMC methods for phylogeny estimation based upon maximum likelihood (Russell and Huelsenbeck), approximation algorithms for maximum parsimony (Rao, Karp, and Papadimitriou), and algorithms for gene family evolution or genomic alignment (Myers and Karp). The UT and UNM researchers will develop (1) new approaches for solving hard optimization problems based upon novel symbolic representations of trees and data (Hunt), (2) new approaches for detecting and reconstructing reticulate evolution (Warnow, Moret, and Linder), (3) divide-and conquer methods for both maximum parsimony and maximum likelihood (Moret, Warnow, and Williams), and (4) methods for reconstructing phylogenies from gene order and content data (Moret, Bader, Warnow, and Jansen).

Simulations: The simulations group will: (1) implement efficient tree comparison algorithms (Kim, Kannan, and Warnow); (2) develop statistical characterization of tree comparison metrics, large deviation theory for tree metrics, and study scaling properties of tree comparison metrics (Kim, Kannan); (3) deliver a small-scale (_100,000 branches) dataset of key molecules (Muse, Hillis); (4) develop models of macro-evolution (Muse, Hillis, Kim); (5) curate available data on well-established phylogenies (Turner); (6) develop protocols for standard benchmark for phylogeny algorithms (Kim, Kannan, Moret, Warnow); (7) develop protocols for scalability benchmark (Kim, Kannan, Moret,Warnow); (8) statistically characterize whole genome evolution (Kim, Meyers); and (9) initiate work on whole genome simulations (Kim, Meyers).
Software Design: Software developers will start replacing old code with new modules (a process to be completed in Year 3). Swofford and Lewis will continue development of phorest, in a version that is immediately compatible with the IT resource, yet also usable in standalone mode. A new stunt run will be conducted, with improved simulated data and new modules; based on lessons from this new run, plans for application and platform scalability will be developed.

Databases: Miranker will continue to lead research into novel database structures and query mechanisms. Piel and the SDSC team will complete the transfer of TreeBase data into the IT resource. Piel, Donoghue, and Mishler will devise a plan for curation of the database over the years; working with SDSC staff, they will also prepare a plan for taking advantage of federation of the new IT resource into the SDSC bioinformatics system.

Outreach: The Berkeley weekend program will be offered regularly and continuously improved through the year from feedback and testing results; plans to include use of a beta-version of the packaged software will be developed. The Lehman summer program will continue. The web pages for the project will be enhanced to reach beyond researchers to a more general college audience. REU supplements will be requested to support more undergraduates through the summer at the various institutions.

SDSC: The SDSC team will develop a beta release of the interface, plan adjustments to hardware and software based on accumulated experience, construct specifications for workflow integration, and identify suitable visualization tools to include in the interface. It will develop and release a first set of standards for producing solutions modules that can be integrated directly into the resource. It will implement the full functionality of TreeBase within the growing database, complete importation of its data, and start federation with SDSC's unified database system.

Year 3

Algorithms: The most promising methods will be extensively evaluated on a variety of tasks. The results will be presented in a manner that biologists as well as other researchers can easily understand the context in which each method is useful. The best of these will be initially integrated with the software efforts at SDSC. The Berkeley and UCSD researchers' focus will be on MCMC methods for maximum likelihood (Russell and Huelsenbeck), gene family evolution and genomic alignment (Myers and Karp), and also lower bounds for maximum parsimony (Rao and Papadimitriou). The UT and UNM group focus will be on heuristics for maximum parsimony and maximum likelihood (Hunt,Warnow, Moret, Bader, and Williams).

Simulations: The group will (1) continue developing statistical methods for tree comparison (Kim, Kannan) ; (2) continue curation of well-established phylogenies (Turner); (3) initiate production work on large-scale simulation of key molecules (Muse, Hillis); (4) develop database protocols for accessing simulated datasets (Kim, Miranker); (5) continue work on whole-genome simulations (Kim, Meyers); and (6) continue work on establishing standard and scalability benchmarks (Kim, Kannan, Moret, Warnow).

Software Design: The authors of various packages will mostly complete the replacement of the first-year "wrapped" versions with new modules written specifically for the IT resource. Swofford and EAC member Lisa Vawter will lead an evaluation effort by outside users—the user panel headed by Vawter and academic researchers. Findings will be compiled and analyzed to develop a plan for revisions. Bader, Berman, and Moret will lead an effort to make all of the code base runnable efficiently on large machines, including clusters of SMPs; they will prepare and submit a proposal for supplementary funding to develop Grid-enabling techniques for the more difficult problems in the code base. All software designers will collaborate closely with the algorithms group in conducting implementation, testing, and refinements of the best algorithms devised to date.

Databases: The research group will consult closely with the SDSC team to integrate its findings into the next stage of DB development. Piel, Donoghue, and Mishler will test the federation system and devise new ways to exploit it in phylogenetic analysis; they will work with Miranker and the SDSC team to design the schema for the next stages of development. Piel will work closely with ATOL partners to include their data and findings within the IT resource.

Outreach: Berkeley weekends events will integrate the alpha release of prior year into their materials, with canned and interactive short demos. Mishler will lead their packaging for use at other team institutions. AMNH and Yale/Peabody will start work on their components—AMNH will produce draft educational modules this year and Peabody will complete much of the high-level design of its museum exhibit. The Lehman program will continue. David Maddison will lead an effort to integrate the project's web site and his Tree of Life web site, with emphasis on access for non-technical visitors. An REU site proposal will be prepared and submitted to ensure continued support of undergraduates for many years.

SDSC: The SDSC team will release a full version of the interface (version 2), complete schema development for the next stage of database development, implement the first version of workflow integration, and develop specifications for the next version of the interface and for data storage and sharing tools.

Year 4

Algorithms: The implementations of the algorithms at SDSC will be further optimized for speed and scalability as well as performance. At this stage the software will be made available to both to biologists working on phylogenies, and to other researchers developing phylogenetic methods.

Simulations: The group will deliver: large-scale key molecule simulated data with database interface, curated empirical dataset of well-established phylogenies, whole-genome simulated dataset, and benchmark suite for statistical accuracy, computational efficiency, and scalability.

Software Design: Swofford and Maddison will lead a focused effort on post-tree analysis software; Swofford will oversee the implementation of the recommendations from the user panel and experimental findings of Year 3. Large scale testing on large datasets produced by the simulation team will be conducted at many of our institutions, led by Moret, Swofford, and Warnow. Throughout the year, will work closely with ATOL partners, helping to analyze their data, learning from the problems exposed in that process.

Databases: Curation will continue as more and more ATOL data are included in the resource. The database team will design and implement a plan to handle the large collections of alternate trees produced on large datasets, using compact representation schemes and sampling techniques.

Outreach: AMNH will field-test its educational modules in local high schools Peabody will complete the design and much of the implementation of its exhibit, including most of its coding. Several institutions will field-test the weekend program developed at Berkeley. The project web site will be enhanced with the addition of a "general public" portal and a portal for children, both in preliminary form.

SDSC: The SDSC team will release a beta version of version 3.0 of the interface, featuring workflow integration, cross-disciplinary query tools (through the federated biological databases at SDSC), and a rich set of reconstruction, evaluation, and analysis tools. It will implement the extended database schema and the planned data storage and sharing tools and integrate the chosen visualization tool(s). It will evaluate the likely status of the IT resource by end of Year 5.

Year 5

Algorithms: The group will advise SDSC on the maintenance and modest improvement to the software that embodies the algorithms that were designed during the course of this project. Publications on the best performing methods are likely to appear in journal form by this time.

Simulations: Publications on the simulation effort should appear in press by this time.

Software Design: The software team will evaluate the progress to date and identify new development targets. It will work closely with the SDSC team to produce a "final" (only with respect to this funding phase) package that will include all of our work and it will run tests with this package on a large variety of platforms. It will carefully document and report the rate of software development over the years of the project and its success in attracting outside contributors, with a view to improving the efficiency of large open-source software development projects.

Databases: The database team will go through the same exercise as the software team: what should be done next and how best to enable other to use what is already present?

Outreach: AMNH will publish its instructional modules; so will Berkeley. Peabody will complete and deploy its exhibit; it will also communicate to other museums around the world how such an exhibit can be assembled, with the intent of duplicating it in many cities. The project web site will get a complete portal for children and will be set up for easy mirroring.

SDSC: The SDSC team will release a full version of the interface (version 3), which will include all tools produced under this funding cycle, as well as a full version of the extended database implementation, with data storage and sharing tools. It will prepare the transition to a new mandate and funding scheme.